Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-430-AC2, 2021

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Resampling and ensemble techniques for improving ANN-based high streamflow forecast accuracy" *by* Everett Snieder et al.

Everett Snieder et al.

usman.khan@lassonde.yorku.ca

Received and published: 1 January 2021

RC2-General.The authors study the effect of resampling techniques, when integrated with ensemble learning frameworks, on the ability of the ANN based regression ensemble learners to improve prediction of high steam flow events. Two case studies are presented, with different temporal resolution and, essentially, hydrologic topology. One individual learner, that is MLP-ANN, is utilized in this study along with two ensemble models (Bagging and Boosting) as well as a randomized set of members (i.e. RWB model). Three resampling plans are examined, RUS, ROS, and SMOTER, to serve as the preprocessing re-sampler stage for the ensemble models. a combination of the

latter is used with the ensemble models and all configurations are evaluated.

This paper attempts to answer an important question which is usually overlooked; can we diminish the heteroscedastic nature of stream flow predictions, which is inevitable when dealing with limited intelligence about the system dynamics (drivers to the instantaneous change in streamflow). The authors are concerned with the most volatile aspect in this setting, high flows, and whether more information can be utilized from the available descriptors' database to alleviate the problem. In general, ensemble learning is one of the few state-of-the-art solutions for improved short- to mid-term streamflow prediction, as individual nonlinear models are inherently instable and conventional statistical approaches sacrifices accuracy for probabilistic interpretability. As the diversity-in-learning mechanism, promoted differently by each ensemble architecture, is assumed to be the major reason to ensemble's generalization ability, resampling techniques are a major interest here. Consequently, it makes sense to study variations of ensemble learning frameworks with respect to the utilized resampling approach. This topic is increasingly gaining attention in the recent, and highly evolving, applied ML community in the field.

The paper is well-written (the chronologic format and section types are not similar to that I am used to, though). The presented results are critical, valid, and to-the-point. The resampling methods are described in organized wording and supported with references. The discussion covers most of the important aspects of this research. To this extent, I believe this paper meets HESS standards and scope, and is worthy of publication, after few important additions and modifications are implemented. Please refer below to the major and minor comments for consideration by the authors.

AC2-General. Thank you for the positive comments on our manuscript and for highlighting the need for the present research. We have addressed each of the major and minor comments below.

RC2-1. The summarization of the individual model calibration is good enough, as the

focus is on the resampling-ensemble models. The use of the PC operator to select features from the predetermined bag of lags also makes sense. However, the authors should track the isolated features and report them. Are they used uniformly among all ensembles? Or does the PC based selection changes per ensemble? Could you also elaborate of the significance of the selected features as well (important for semi-physical validation of the used features).

AC2-1. Thank you for the recommendation. The selected input features are now listed in Table A1 of the Appendix (please refer to attached supplement). Since PC is a model free method, the outcome is independent from the training method, thus the feature set is constant for each method. In fact, this is the reason we selected to use the PC method, since it allows a consistent feature set for all model configurations. As for the semi-physical validation of the features, this is discussed in detail in [1]. However, in general, due to the autocorrelated nature of the Bow River, the inputs are dominated by autoregressive input variables, upstream flow, and temperature, which drives snowmelt. Whereas, for the Don River, the inputs are a mixture of precipitation and upstream and/or lagged flows. This to be expected, based on results in [1], as well as the nature of the two watersheds.

RC2-2. Please include pseudo-algorithm table for each resampling plan. This is very important for recreation proposes (the two utilized ensemble models, on the other hand, are well-studied in the broad literature and do not require detailed description; though I would prefer to see a mathematical description of the models to further acquaint the readers with them as ensemble learning is still not common across all fields of HESS).

AC2-2. Thank you for the recommendation. Pseudocode has been provided for the three resampling methods and three ensemble methods (please refer to the attached supplement). Including the code for the ensemble methods (in addition to the resampling plans) is also relevant, as the resampling methods are embedded within ensemble methods.Thus, both have been included.

RC2-3. Also, please elaborate more on the distinction between RUS and ROS (for a first glance, the wording makes RUS looks like a special case of ROS, but they are very different in reality and have distinct effect on the model performance). Please elaborate more on your choice of the ROS configuration, and why not present an array of results related to the OSR (ratio vs. performance for example).

You are correct: ROS and RUS are distinct (i.e., RUS is not a special case of ROS). As stated in-text, RUS undersamples data as to achieve a balanced training set with no duplicates whereas ROS includes all available data and creates duplicates. Some studies have compared the two sampling techniques, however neither one of these two methods consistently outperforms the other, thus both are included in our study [2-4]. We have now included the pseudo-algorithms for each (see response AC2-2 above) to help clarify the distinction. All three resampling methods (RUS, ROS, and SMOTER) are configured such that the number of typical and high stage samples in the resampled dataset are equal. In this research, since the 80th percentile stage is used to distinguish between typical and high flows, ROS resamples the high stage subset by 500%.

Furthermore, the effects of SMOTER-based resampling have been explicitly quantified by calculating the ratio between SMOTER-Bagging and Bagging in two new figures. Firstly, the effects of SMOTER resampling is assessed across the high stage threshold value, thus rate of oversampling. Next, the resampling effects are analysed across the number of ensemble members, thus the number of resampling repetitions. Please refer to Figures 11 and 12, and associated text in the attached supplement.

RC2-4. Please modify the ensemble learning section to have a more concise summary of ensemble learning, diversity-in-learning concept and the effect of resampling as part of the latter. Please provide references from the pure literature.

AC2-4. We apologise for the verbose section on ensemble learning and thank you for this recommendation. Some of the oversimple background on ensemble methods have

been trimmed from the text. Relevant background on ensemble learning and sources of diversity have been added to the text and collection of cited works have been expanded: “Ensembles are collections of models, each trained on different subsets of the available training data and combined to form discrete ensemble prediction (Alobaidi et al., 2019). It is well established that ensemble-based methods improve model stability and generalisability (Alobaidi et al., 2019; Brown et al., 2005). Recent advances in ensemble learning have emphasised the importance of diversity-in-learning (Alobaidi et al., 2019). Diversity can be generated both implicitly and explicitly through a variety of methods, some of which include varying the initial set of model parameters, varying the model topology, varying the training algorithm, and varying the training data (Sharkey, 1996; Brown et al., 2005). The largest source of diversity in the ensembles under study is attributable with varying the training data, which occurs both in the various resampling methods described above and the in some cases, the ensemble algorithms. Only homogeneous ensembles are used in this work, thus no diversity is obtained through varying the model topology or training algorithm (Zhang et al., 2018; Alobaidi et al., 2019). Ensemble predictions are combined to form a single discrete prediction. Ensembles that are combined to produce discrete predictions have been proven to outperform single models by reducing model bias and variance, thus improving overall model generalisability (Brown et al., 2005; Sharkey, 1996; Shu and Burn, 2004; Alobaidi et al., 2019). This has contributed to their widespread application in hydrological modelling (Abrahart et al., 2012). In some cases, ensembles are not combined, and the collection of predictions are used to estimate the uncertainty associated with the diversity between ensemble members (Tiwari and Chatterjee, 2010; Abrahart et al., 2012). While this approach has obvious advantages, it is not possible for all types of ensembles, such as the boosting methods used in this work.”

RC2-5. It is important to note that the RWB model, contrast to what has been suggested in the applied literature, is NOT an ensemble model. Ensemble learning has three stages one of which is the resampling of the available intelligence. RWB violates this and should not be considered an ensemble for the sake of clarity. However, the

randomization of weights within the individual learners are a major source of diversity, as shown in the literature, which promotes similar behaviour improvement in this model as other ensembles. RWB can be considered as middle ground between individual and ensemble learning. Please fix this issue.

AC2-5. Thank you for clarifying this point. While we are not aware of an agreed-upon definition for what qualifies an ensemble model in literature [5,6]. Quoting from [6]: “This is perhaps the most common way of generating an ensemble, but is now generally accepted as the least effective method of achieving good diversity; many authors use this as a default benchmark for their own methods” [6].

You are correct that RWB is unlike the other ensemble methods in that it does not have a source of diversity-in-learning. We have kept the organisation of the manuscript the same, but added text to distinguish RWB from the ensemble learning methods in this regard. We have adjusted the RWB text accordingly: “While not technically a form of ensemble learning, repeatedly randomising the weights and biases of ANNs is one of the simplest and most common methods for achieving diversity among a collection of models, thus it acts as a good comparison point for the proceeding ensemble methods (Brown et al., 2005). In this method, members are only distinguished by the randomisation of the initial parameter values (i.e., the initial weights and biases for ANNs in this research) used for training.”

RC2-6. I do not think that the authors replace the ensemble’s native resampling technique with the suggested approaches, but rather add them as a preprocessing phase of the available data, as indicated in the manuscript (also, it would be impossible to replace the resampling approach in boosting!). Hence, it seems like there is “double resampling” which occurs in the modified ensemble models. this is interesting (as shown in the obtained results). Please elaborate on the underlying effect of the ensemble’s native resampling technique on the preprocessing one. For example, Bagging’s uniform resampling plan has little effect, while that of Boosting has a very strong effect (I am thinking it nullifies the preprocessing resampler!).

AC2-6. Thank you for the recommendation. We do, in fact, imbed the resampling methods within the ensemble algorithms. In machine learning literature this has been referred to as hybridisation [7]. For example, the hybridisation of oversampling with bagging achieves the benefits of increased representation of infrequent values (high flows) and the diversity obtained through repeated resampling with replacement [7]. The most complex of these combinations, AdaBoost with SMOTE, has been demonstrated for classification [8] and regression [9] in existing literature. AdaBoost is configured with reweighting the ANN cost function, so that there is no double resampling; however, such a resampling scheme is possible. We believe that our proposed method, for determining sample weights for synthetic samples in this algorithm, is novel and an improvement over existing implementations of SMOTE with AdaBoost. The pseudocode produced in response to RC2-2 elaborates on the implementation of these methods.

RC2-7. Linking the previous comment with RWB, having a preprocessing resampler makes RWB a true ensemble model here. More importantly, the double-resampling effect is absent here, making it a great opportunity to elaborate on the difference between modified RWB and other ensembles! Please do so.

AC2-7. As stated in our response to RC2-6, the resampling methods are embedded within the ensemble methods. Subsequently, for RWB, resampling occurs as a one-time preprocessing step and the series of models with randomised weights and biases are trained using the fixed, resampled data. The ensemble methods do not have double resampling, as the resampling methods are not used in preprocessing. Please refer to the pseudocodes above (AC2-2) for a description of the methods.

RC2-8. Figures 6 to 9 and Tables 5 and 6 are very important and provide most of the critical information to show how added resampling promotes improved high-flow prediction (and overall prediction also). In the same time, I sincerely think that a major result is missing here, which deals with quantifying the effect of the resampling approaches on the ensembles. I think that the paper requires a figure showing the

change of ensemble performance versus ensemble size. This figure should at least depict the Bagging model and the modified Bagging model. I recommend adding the RWB and Modified RWB. This is very important to cross-examine the change of performance, per ensemble size, between the normal and modified ensembles and provide more insight on the effect of the preprocessing phase.

AC2-8. Thank you for this recommendation. We have included a formalised ensemble size grid-search for ensemble sizes ranging from 2-100 for each of the base ensemble methods (i.e., with no added resampling). Please refer to Fig. 6 and associated text in the attached supplement.

We have also conducted a grid-search of ensemble size to assess the SMOTER resampling effect on the Bagging algorithm. The new figure (Fig. 12) and associated text is included in the attached supplement.

RC2-9. The fact that tables 5 and 6 show Boosting to perform the worst validates the results obtained, as Boosting performance deteriorates in the presence of “hard instances” in general and is more applicable to classification applications (at a larger ensemble size, the obtained combiners’ weights seem to dilute in efficacy when performing regression. But in classification, they are powerful especially in binary classification due to the sign significance rather than magnitude of the weight). The authors attempt to explain the reasons to the deficiency of Boosting ensemble in the paper but I think they can elaborate more.

AC2-9. The authors have added text addressing the overfitting tendencies of the boosted ensembles:

“The overfitting produced by the boosting methods is consistent with previous research, which finds that boosting is sometimes prone to overfitting on real-world datasets (Vezhnevets and Barinova, 2007). One reason that the improvements made by the boosting methods (AdaBoost and LSBoost) are not more substantial may be due to the use of ANNs as individual learners. ANNs typically have more degrees of freedom

compared to the decision trees that are most commonly used as individual learners; thus, the additional complexity offered by boosting does little to improve model predictions. Additionally, the boosting methods further increase the effective degrees of freedom of the predictions. Nevertheless, these methods still tend to improve performance over the base model case. Ensembles of less complex models such as regression trees are expected to produce relatively larger improvements when relative to the single model predictions."

RC2-10. More importantly, the information in comments 3, 6 and 7 should be considered here (when discussing tables 5 and 6). For example in table 5, when considering RWB and Bagging models, why was ROS based models the worst (I think because there are a lot of data), but in the same time SMOTER variations did provide good performance in few of the metrics. Why did SMOTED-RWB have an unusually low PITF but good performance with respect to other metrics? Please provide more details similar to this.

AC2-10. Thank you for this insightful recommendation. We have added specific remarks on the overfitting effects of ROS and the difference in performance and diversity between RUS-RWB and RUS-Bagging. The text is copied below, and this has also been addressed elsewhere in the response to the reviewer's comments:

"ROS often exhibits poorer performance than SMOTER and RUS. Previous research has noted the tendency for ROS-based 510 models to overfit, due to the high number of duplicate samples (Yap et al., 2014). RUS, despite using considerable less training data for each individual learner, is not as prone to overfitting as ROS. The RUS-Bagging models consistently outperform the RUS-RWB models; this may be due to the repeated resampling, thus RUS-Bagging uses much more of the original training samples, while RUS-RWB only uses 20% of the original data."

RC2-11. There are few typos and referencing issues. Please revise the manuscript for this. A few examples are provided below:

Line 4: "compare three resampling;" I think it is missing a word! Table 3: second column. Do you mean "variable" or "feature"? Line 113: please include the term "individual learners". Line 163: comma is missing after the reference. Line 164: the reference seems to be in the middle of the sentence. I noticed you do this often. Please try to minimize this to enable smother flow of the idea. Line 168: "for handling imbalance data. . ." do you mean "imbalanced". Line 170: "that featuring each. . .". Please fix. Please italicize all symbols or feature names, such as N, theta, PI, CE, etc., across the manuscript. Line 205: the reference format needs a fix. Line 255 to 260: please rewrite the ensemble learning summary as discussed in the major comment. Line 262: "Ensembles members are.." please remove the "s". Line 310: do you mean "regional flood quantiles"?.

AC2-11. Thank you for pointing out these typos - we have corrected these issues in the revised version of the manuscript.

References [1] E. Snieder, R. Shakir, and U. T. Khan, "A comprehensive comparison of four input variable selection methods for artificial neural network flow forecasting models," J. Hydrol., vol. 583, p. 124299, Apr. 2020. [2] B. W. Yap, K. A. Rani, H. A. A. Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets," Lect. Notes Electr. Eng., vol. 285 LNEE, pp. 13–22, 2014. [3] M. Bach, A. Werner, J. Åżywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," Inf. Sci. (Ny)., vol. 384, pp. 174–190, Apr. 2017. [4] J. F. Díez-Pastor, J. J. Rodríguez, C. I. García-Osorio, and L. I. Kuncheva, "Diversity techniques improve the performance of the best imbalance learning ensembles," Inf. Sci. (Ny)., vol. 325, pp. 98–117, Dec. 2015. [5] D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," J. Artif. Intell. Res., vol. 11, pp. 169–198, Aug. 1999. [6] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," Inf. Fusion, vol. 6, no. 1, pp. 5–20, 2005. [7] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera,

“A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, vol. 42, no. 4. pp. 463–484, 2012. [8] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, “Random Balance: Ensembles of variable priors classifiers for imbalanced data,” Knowledge-Based Syst., vol. 85, pp. 96–111, Sep. 2015. [9] N. Moniz, R. Ribeiro, V. Cerqueira, and N. Chawla, “SMOTEBoost for Regression: Improving the Prediction of Extreme Values,” in 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 150–159.

Please also note the supplement to this comment:
https://hess.copernicus.org/preprints/hess-2020-430/hess-2020-430-AC2-supplement.pdf

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-430, 2020.